More About Datasets

On the last page, we discussed the main features of *datastores* and *datasets*. Now, let's look a little more closely at how datasets work in Azure Machine Learning.

Key points to remember about datasets:

- They are used to interact with your data in the datastore and to package data into consumable objects.
- They can be created from local files, public URLs, Azure Open Datasets, and files uploaded to the datastores.
- They are not copies of the data but *references* that point to the original data. This means that no extra storage cost is incurred when you create a new dataset.
- Once a dataset is registered in Azure ML workspace, you can share it and reuse it across various other experiments without data ingestion complexities.

In summary, here are some of the main things that datasets allow you to do:

- Have a single copy of some data in your storage, but reference it multiple times—so that you don't need to create multiple copies each time you need that data available.
- Access data during model training without specifying connection strings or data paths.
- More easily share data and collaborate with other users.
- Bookmark the state of your data by using dataset versioning

You would do versioning most typically when:

- New data is available for retraining.
- When you are applying different approaches to data preparation or feature engineering.

Keep in mind that there are two dataset types supported in Azure ML Workspace:

- The **Tabular Dataset**, which represents data in a tabular format created by parsing the provided file or list of files.
- The **Web URL (File Dataset)**, which references single or multiple files in datastores or from public URLs.

We'll get practice working with these dataset types in the upcoming labs.

QUIZ QUESTION

Which of the following are true statements about Azure Machine Learning datasets?

(Select all that apply.)

□ Datasets are copies of your data files that are copied by Azure onto your machine as needed.

✓ Datasets are references that point to the data in your storage service; they are not actually copies, so no extra storage cost is incurred.

✓ If you don't have an Azure storage service, you can create a dataset directly from local files, public urls, or an Azure Open Dataset.

SUBMIT